# Development of Corpus Resources for Empirical Translation Studies

**Meng JI**
University of Western Australia

*From its inception in the 1990s, corpus-based translation studies or CBTS represents a growing area of research which is largely based on the development of digital research resources and technologies. This paper offers an overview of the development of the field in the last twenty years, highlighting the importance of developing pragmatic and versatile analytical techniques in order to optimise use of corpus resources and tools based on current natural language processing technologies. This includes the development of small-scale yet effective language corpora and the devise of annotation schemes and strategies to serve specific research purposes that are termed as problem-oriented corpus annotation here.*

**Keywords: corpus-based translation studies; corpus analysis; problem-oriented annotation**

## 1. Introduction

The rapid development of large-scale parallel or multilingual corpora has greatly advanced the study of translational or multilingual texts and related social and cultural issues. The systematic exploration of newly developed language corpora has given rise to emerging research areas such as corpus stylistics, cognitive stylistics, corpus-based translation studies which are distinctively interdisciplinary and descriptive. Since the inception of corpus translation studies in the 1990s, this empirical branch of translation studies has grown into one of the most promising fields of translation research that is

widely taught in postgraduate courses of Translation Studies worldwide. What lies at the heart of this emerging discipline is the design and construction of increasingly larger language data bases and the development of effective and reliable corpus analytical techniques (Oakes and Ji, 2012).

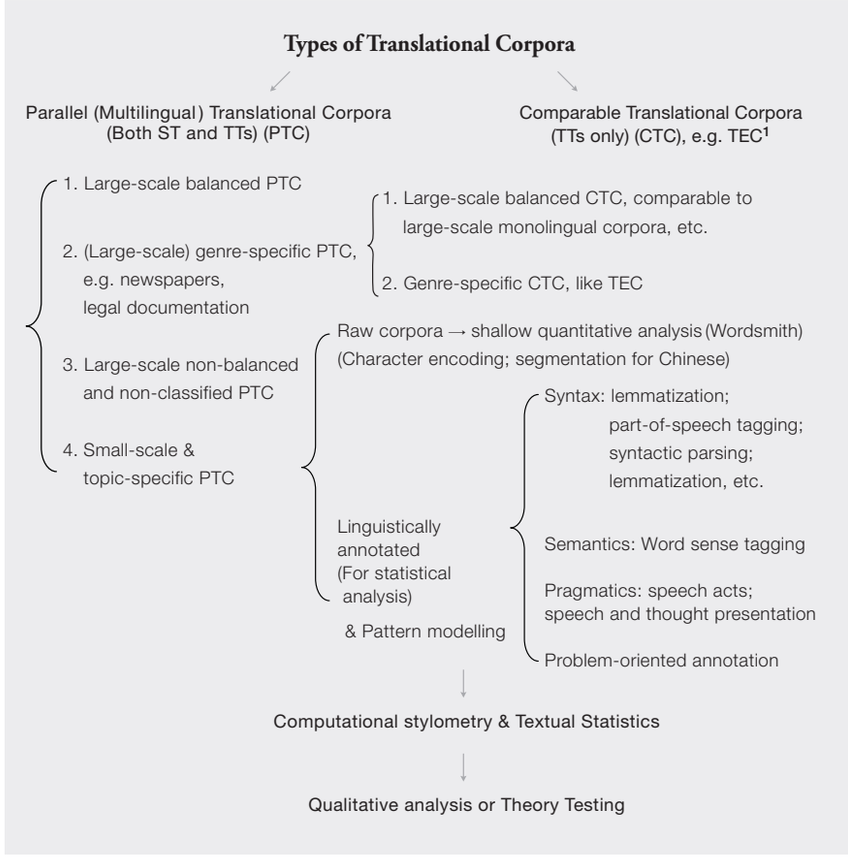
Technical advancement in the development of digital language resources has transformed the study of literature and language since the second half of the twentieth century (Stubbs, 1996; Thomas and Short, 1996; Lawler and Dry, 1998). From its inception in the 1990s, empirical or to be more specific, corpus-based translation studies has always sought to take fully advantage of large-scale digital resources to advance the aims and purposes of translation research (Baker, 1995: 223-243). Baker's thesis is a milestone in the development of corpus Translation Studies as an independent research paradigm. The investigation of quantitative linguistic data collected in language corpora by using computational techniques has fundamentally changed the way we observe, analyse and conceptualise translation.

From early efforts and discussions on the construction of parallel or comparable corpora of translation (McEnery and Wilson, 1993; Teubert, 1996), electronic resources created for Translation Studies range from small-scale topic-specific corpora to statistically-built parallel corpora. An important feature of annotated language corpora is the rich linguistic information supplemented to raw corpus texts by using automatic tagging systems which have become increasingly fine-grained and of high precision. The significance of creating linguistically rich language corpora is that marked databases make important preparation for the identification and retrieval of textual patterns which form the basis of the formulation and verification of theoretical hypotheses.

Figure 1 presents important types of language corpora which are relevant to empirical translation studies. While some of the corpora are widely available for research purposes, others are still under development either due to technical issues or the lack of necessary translational data. The scale and diversity of translational corpora that have been developed in the last twenty years or so are rather conspicuous. The advances made in parallel corpus construction would seem more prominent, if we take into account the much higher levels of difficulty implied in solving technical problems relating to parallel text matching and alignment, especially working with typologically different languages (Piao, 2002).

At the first tier, there are two major types of translational corpora: parallel

Figure 1. Types of Translational Corpora




translational corpora (PTC) and comparable translational corpora (CTC). There has been some confusion in the literature regarding the establishment of a consistent terminological framework for corpus type categorization (Baker, 1999; Hunston, 2002). The differences between the two may be better described and understood by looking at their underlying structural features: while PTC contains both the source and target texts, CTC is a compilation of

[1] Translational English Corpora http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm

translated texts only, with a view to investigating the nature and regularities of translated language (Baker, 1995: 223-43).

Within each category, PTC or CTC can be further classified according to the text types that they may cover, whether they are large-scale balanced corpora (LSBC) or genre-specific corpora (GSC). The purpose of building large-scale balanced corpora is to investigate general linguistic features of the language in use; whereas the compilation of genre-specific translational corpora aims primarily to address research questions regarding specific aspects of translated language or within certain text domains.

Both types of corpora, i.e. LSBC and GSC, may be explored in the construction of a translational corpus platform; however, experience shows that unlike the construction of monolingual corpora (one language only), large-scale national corpora like the British National Corpus (BNC), the Modern Chinese Language Corpus (MCLC), the American National Corpus (ANC), the Korean National Corpus (KNC), etc., the development of LSBC, which are populated by translated texts only, is both size-limited in size and much less balanced. It is not difficult to understand why this should happen: while native-speakers' use of language is very frequent in every aspect of social life, translated language may be found only in a rather limited range of communicative environments, such as language teaching and learning, conference interpretation, and mainly, published translations.

The obvious shortage of translation material makes the establishment of a balanced framework of sampling notably difficult, which in turn has restricted the scope of contrastive studies, mainly between translated and non-translated languages, that current versions of PTC and CTC may be able to cover. Most PTC and CTC developed so far are genre-specific corpora, which favour a restricted range of text genres that may be easily collected from published translations, e.g. legal texts, official documents, fictions, commercial fliers or technical manuals, etc.

Automatic construction of parallel/multilingual translational corpora usually approaches technical issues quite differently from the way it is done in the compilation of a comparable translational corpus. In building up a CTC, the establishment of a well balanced sampling framework may be crucial; without an adequate text selection scheme, the comparison between corpora in translated languages and original languages – in this case, translated language is seen as a legitimate language type in its own right, which is precisely the rationale behind the construction of TEC –, will be seriously compromised.



On the other hand, for the construction of a PTC, the first technical snag is a proper automatic alignment of source and target texts at a word or sentential level, which represents one of the most difficult problems in natural language processing and machine translation (Véronis, 2000: 25).

Automatic parallel corpus construction, which started with Gale and Church's famous sentence-length principle (Gale and Church, 1991), has evolved rapidly into a rich theoretical paradigm, which attempts to integrate both empirical linguistic cues, such as similar sentence length or cognate character words contained in the ST or its corresponding TT segments, and statistical alignment techniques, e.g. sentence-based dynamic programming (Gale and Church, 1991; Brown et al. 1991) and language-independent fuzzy chunk matching (Deng et al. 2006).

The continuous and collaborative efforts made by computational linguists have greatly improved the efficiency of automatic parallel text alignment systems, as well as facilitated the construction of a number of domain-specific parallel corpora, prominently in intra-European languages and other well explored language pairs, such as English-Chinese or English-Japanese.

## 2. Pragmatic Use of Small-Scale Corpora to Address Specific Research Questions

The enormity of costs and copyright issues involved in the development of large-scale parallel corpora often prevent them from becoming freely accessible to translation scholars or research students. This in turn has given rise to another type of popular parallel corpora shown in Figure 1, i.e. small-scale topic-specific PTC, which in fact constitute the mainstream corpora used today in most individual research projects in line with corpus-based translation studies. As the most frequently used type of corpora in translation research, small-scale topic-specific parallel corpora are usually built to a large extent manually with specific questions born in the researcher's mind.

To start with, small-scale DIY corpora normally require a great deal of effort and dedication on the part of the researcher to prepare the textual material especially selected for the purpose, either inspired by certain methodological orientations (for normalization see Kenny, 2001; Mundy, 1998; for explicitation, see Pápai, 2004; Purrtinen, 2004; for simplification, see Laviosa, 1997 and 2000), or by adopting an openly corpus-driven approach to the

particular ST/TT or TT/TT pairs under investigation (Saldanha, 2005).

Due to the limited size of small-scale and topic-specific corpora, the specific questions that the researcher may be seeking to answer are exploratory in nature. That is to say, the findings obtained as an end-result of each individual project are largely circumscribed and are usually domain-or author-specific. Nevertheless, the usefulness of this type of DIY corpora, when studied in conjunction with large scale comparable corpora, non-translational or translational, may be maximally extended; and the findings revealed in those tailor-made corpora may also be further contrasted with the data collected from more general referential corpora.

Kenny (2001) studies the translational features of normalization in English translations of German fictions. It exemplifies the development and exploration of purposely constructed genre-specific translation corpora. Normalisation is part of a set of hypothesised translational features that are proposed as generally existent in translated languages, despite the language pairs involved. These are known as translation universals and norms which have provided the focus for a large number of case studies pursued in line with corpus translation studies. Normalisation is understood as a tendency in translation to exaggerate features of the target language and conform to its typical linguistic patterns. Normalisation may be detected at various levels including syntactic, lexical and grammatical.

At the lexical level, normalization is shown in the use of more conventional lexical items and more conventional ways of combining lexical items than non-translated source text language. For the purpose of her research, Kenny constructs a two-million Germany-English parallel corpus of literary texts. Frequency-ranked wordlists are used to identify potentially creative hapax legomena in source texts; the creative status of such hapax legomena is then verified using standard lexicographical sources, native speaker judgments, and, most importantly, a reference corpus of non-translated German texts.

The study finds that while around 44% of creative hapax legomena identified in the German source texts are normalized in their English translations. A number of factors, of both a textual-linguistic and a demographic nature that may condition normalization are proposed. However, given the small number of examples studied in each relevant category, Kenny warns that any conclusions are necessarily tentative and await verification in future, scaled-up studies. In fact, the technical issue highlighted in Kenny (2001), i.e. the lack of sufficient corpus data to verify a hypothesis may be effectively solved by



combining the use of genre or topic-specific corpora with large-scale balanced monolingual and/or translational corpora; or comparing two large-scale language corpora of similar sampling structure.

Ji (2010) represents another important approach to the study of translation corpora. It focuses on the comparison of different translations of the same source text in an effort to identify, analyse and explain the systematic differences among translations. In Figure 2, the Chinese translation shown on the left is the translation by Yang in 1978. The result of the corpus analysis is then interpreted in light of translation stylistics. For the purpose of her research, Ji constructed a parallel Castilian-Chinese corpus.

The parallel corpus constructed includes modern Chinese versions of Miguel Cervantes' *Don Quijote de La Mancha* written in seventeenth century Castilian. The two Chinese translations selected for comparison were by Yang Jiang in 1978 and by Liu Jingsheng in 1995. Through the corpus-based analysis of the two corpora, it is found that when compared to the earlier Chinese version (Yang, 1978) of the Spanish novel, the use of Chinese idioms in the latter translation (Liu, 1995) has been greatly enhanced.

The corpus analysis seem to suggest that the enhanced use of Chinese idioms is an important stylistic feature of Liu's translation when compared to Yang's version. However, it is suspected that given the gap between the two

Figure 2. A parallel Castilian-Chinese corpus of Don Quijote de La Mancha (Part 1)

| | |
|---|---|
| 为了狠狠地踢一顿叛徒加拉隆，他情愿献出他的女管家，甚至可以再赔上他的外甥女 | Diera él, por dar una mano de coces al traidor de Galalón, al ama que tenía, y aun a su sobrina de añadidura. |
| 实际上，他理性已尽失。他产生了一个世界上所有疯子都不曾想过的怪诞想法，自己倒认为既合适又有必要，既可以提高自己的声望，还可以报效他的国家。他要做个游侠骑士，带着他的甲胄和马走遍世界，八方征险，实施他在小说里看到的游侠骑士所做的一切，赴汤蹈火，报尽天下仇，而后留芳千古。 | En efeto, rematado ya su juicio, vino a dar en el más extraño pensamiento que jamás dio loco en el mundo; y fue que le pareció convenible y necesario, así para el aumento de su honra como para el servicio de su república, hacerse caballero andante, y irse por todo el mundo con sus armas y caballo a buscar las aventuras y a ejercitarse en todo aquello que él había leído que los caballeros andantes se ejercitaban, deshaciendo todo género de agravio, y poniéndose en ocasiones y peligros donde, acabándolos, cobrase eterno nombre y fama. |

| | |
|---|---|
| 可怜的他已经在想象靠自己双臂的力量，起码得统治特拉彼松达帝国。想到这些,他心中陶然,而且从中体验到了一种奇特的快感，于是他立即将愿望付诸行动。 | Imaginábase el pobre ya coronado por el valor de su brazo, por lo menos, del imperio de Trapisonda; y así, con estos tan agradables pensamientos, llevado del extraño gusto que en ellos sentía, se dio priesa a poner en efeto lo que deseaba. |
| 他首先做的就是清洗他的曾祖父留下的甲胄。甲胄长年不用，被遗忘在一个角落里,已经生锈发霉。 | Y lo primero que hizo fue limpiar unas armas que habían sido de sus bisabuelos, que, tomadas de orín y llenas de moho, luengos siglos había que estaban puestas y olvidadas en un rincón. |
| 他把甲胃洗干净，尽可能地拾掇好，可是他发现了一个大毛病，就是没有完整的头盔，只有简单的顶盔。 | Limpiólas y aderezólas lo mejor que pudo, pero vio que tenían una gran falta, y era que no tenían celada de encaje, sino morrión simple; |
| 不过，他可以设法补救。他用纸壳做了半个头盔接在顶盔上，看起来像个完整的头盔。 | mas a esto suplió su industria, porque de cartones hizo un modo de media celada, que, encajada con el morrión, hacían una apariencia de celada entera. |
| 为了试试头盔是否结实，是否能够抵御刀击，他拔剑扎了两下。结果，刚在一个地方扎了一下，他一星期的成果就毁坏了，看到这么容易就把它弄碎了，他颇感不快，他又做了一个头盔。为了保证头盔不会再次被毁坏，他在里面装了几根铁棍。他对自己的头盔感到满意，不愿意再做试验，就当它是个完美的头盔。 | Es verdad que para probar si era fuerte y podía estar al riesgo de una cuchillada, sacó su espada y le dio dos golpes, y con el primero y en un punto deshizo lo que había hecho en una semana; y no dejó de parecerle mal la facilidad con que la había hecho pedazos, y, por asegurarse deste peligro, la tornó a hacer de nuevo, poniéndole unas barras de hierro por de dentro, de tal manera, que él quedó satisfecho de su fortaleza y, sin querer hacer nueva experiencia della, la diputó y tuvo por celada finísima de encaje. |
| 然后，他去看马。虽然那马的蹄裂好比一个雷阿尔，毛病比戈内拉那匹皮包骨头的马毛病还多，他还是觉得，无论压力山大的骏马布塞法洛还是熙德的骏马巴别卡，都不能与之相比。 | Fue luego a ver su rocín, y aunque tenía más cuartos que un real y más tachas que el caballo de Gonela, que tantum pellis et ossa fuit, le pareció que ni el Bucéfalo de Alejandro ni Babieca el del Cid con él se igualaban. |
| 他用了四天时间给马起名。因为（据他自言自语），像他那样有名望、心地善良的骑士的马没有个赫赫大名就太不像话了。他要给马起个名字，让人知道，在他成为游侠之前他的声明，后果又怎么样。 | Cuatro días se le pasaron en imaginar qué nombre le pondría; porque (según se decía él a sí mismo) no era razón que caballo de caballero tan famoso, y tan bueno él por sí, estuviese sin nombre conocido; y ansí, procuraba acomodársele de manera que declarase quién había sido antes que fuese de caballero andante, y lo que era entonces; |



| | |
|---|---|
| 主人地位变，马名随之改，这也是合情合理的。得起个鼎鼎显赫、如雷贯耳的名字，才能与他的新品第、新行当相匹配。 | que declarase quién había sido antes que fuese de caballero andante, y lo que era entonces; pues estaba muy puesto en razón que, mudando su señor estado, mudase él también el nombre, y le cobrase famoso y de estruendo, como convenía a la nueva orden y al nuevo ejercicio que ya profesaba; |
| 他造了很多名字，都不行，再补充，又去掉。最后，凭记忆加想象，才选定叫罗西南多。他觉得这个名字高雅、响亮，表示在此之前，它是一匹瘦马，而今却在世界上首屈一指。 | y así, después de muchos nombres que formó, borró y quitó, añadió, deshizo y tornó a hacer en su memoria e imaginación, al fin le vino a llamar Rocinante, nombre, a su parecer, alto, sonoro y significativo de lo que había sido cuando fue rocín, antes de lo que ahora era, que era antes y primero de todos los rocines del mundo. |
| 给马起了个称心如意的名字之后，他又想给自己起个名字。这又想了八天，最后才想起叫唐吉坷德。前面谈到，这个真实故事的作者认为他肯定叫基哈达，而不是像别人说的那样叫克萨达。 | Puesto nombre, y tan a su gusto, a su caballo, quiso ponérsele a sí mismo, y en este pensamiento duró otros ocho días, y al cabo se vino a llamar don Quijote; de donde, como queda dicho, tomaron ocasión los autores desta tan verdadera historia que, sin duda, se debía de llamar Quijada, y no Quesada, como otros quisieron decir. |
| 不过，想到勇敢的阿马迪斯不满足于叫阿马迪斯，还要把王国和家乡的名字加上，为故里增光，叫高卢的阿马迪斯，这位优秀的骑士也想把老家的名字加在自己的名字上，就叫曼查的唐吉诃德。他觉得这样既可以表明自己的籍贯，还可以为故乡带来荣耀。 | Pero, acordándose que el valeroso Amadís no sólo se había contentado con llamarse Amadís a secas, sino que añadió el nombre de su reino y patria, por hacerla famosa, y se llamó Amadís de Gaula, así quiso, como buen caballero, añadir al suyo el nombre de la suya y llamarse don Quijote de la Mancha, con que, a su parecer, declaraba muy al vivo su linaje y patria, y la honraba con tomar el sobrenombre della. |
| 洗净了甲胄，把顶盔做成了头盔，又为马和自己起了名字，他想，就差一个恋人了。没有爱情的游侠骑士就好像一棵树无叶无果，一个躯体没有灵魂。 | Limpias, pues, sus armas, hecho del morrión celada, puesto nombre a su rocín y confirmándose a sí mismo, se dio a entender que no le faltaba otra cosa sino buscar una dama de quien enamorarse: porque el caballero andante sin amores era árbol sin hojas y sin fruto y cuerpo sin alma. |
| 他自语道：“假如我倒霉或走运，在什么地方碰到某个巨人，这对游侠骑士是常有的事，我就一下子把他打翻在地或拦腰斩断，或者最终把他战胜，降伏了他。 | Decíase él: Si yo, por malos de mis pecados, o por mi buena suerte, me encuentro por ahí con algún gigante, como de ordinario les acontece a los caballeros andantes, y le derribo de un |

Table 1. Distribution of idioms in LCMC versus UCLA Chinese corpus

| Code | Text Type | Raw Frequency (LCMC) | Raw Frequency (UCLA) |
|---|---|---|---|
| AD | Adventure/Martial Arts Fiction | 338 | 300 |
| ES | Essays and Biographies | 931 | 363 |
| GF | General Fiction | 290 | 223 |
| HU | Humor | 108 | 76 |
| MY | Mystery/Detective Fiction | 266 | 493 |
| NED | News Editorials | 369 | 111 |
| NREP | News Reportage | 484 | 236 |
| NREV | News Reviews | 249 | 117 |
| PL | Popular Lore | 501 | 171 |
| RE | Religion | 112 | 7 |
| REP | Reports/Official Documents | 108 | 36 |
| RO | Romantic Fiction | 378 | 263 |
| SC | Science (Academic Prose) | 344 | 51 |
| SF | Science Fiction | 45 | 255 |
| SK | Skills/Trades/Hobbies | 244 | 9 |
| **Total** | **Total** | 4767 | 2711 |

translations produced in the 1970s and the 1990s, respectively, the general variation and change of modern Chinese lexis may well be a contextual factor that explains the differences between Yang's and Liu's translation. To address this pending question, Ji investigates the evolving nature of Mandarin Chinese through a comparative study of the distribution of Chinese idioms in two large-scale modern Chinese monolingual corpora, i.e. Lancaster Corpus of Mandarin Chinese, also known as LCMC (1990s) and the UCLA Chinese Corpus (early 2000s).

These two corpora have been constructed by following the same sampling framework as that of the Brown or the LOB corpus, and are thus essentially comparable. The result of the corpus comparison shows that idioms, which represent the most conventionalized part of Chinese, seem to have undergone a considerable change in the last decade of the twentieth century, for when compared to the LCMC, many of the text types or genres have witnessed a noticeable decrease in the occurrence of idioms in the UCLA corpus.

As two widely distributed monolingual corpora of modern Chinese, both LCMC and UCLA Corpus have been built to address the increasing need for large-scale comparable corpora to do contrastive language studies, usually in combination with purposely-built specific corpora of much smaller size. A quantitative study of the two diachronically successive corpora brings valuable insights into the changing nature of Chinese, as being focused upon at a particular historical point. The linguistic phenomenon under investigation is the distribution of Chinese idioms, as a core part of the language, among the various text types included in the two corpora, which add up to some fifteen categories.

Table 1 exhibits the raw frequency of idioms in different text genres, which is an initial comparison of the two monolingual corpora of Chinese. It should be noted that the first impression that we may have of such comparison may turn out to be misleading, due to the different size of the two corpora: while the LCMC contains one million tokens[2], the current version of the UCLA corpus holds 687, 634 running words in its collection[3]. As a result, it would be rather difficult to tell from the outset whether the two corpora genuinely differ from each other with regards to the distribution of idioms across the fifteen text types. The statistical procedure Pearson's correlation test has been employed, which yields the important statistical result shown in Table 2.

Pearson's correlation test is widely used in corpus linguistics to test the strength of association between different corpus texts. It does not assume any causal relationship between the variables under test and may only deal with continuous data. It expresses the strength of correlation numerically through the correlation coefficient, R, which varies from menus one to one as the maximum values at two extremes. Table 2 shows that firstly, the mean frequency of idioms in the LCMC is as high as 317.8, which is almost twice that of the UCLA corpus. The computed coefficient of the correlation model is approximately 0.435, whose further interpretation requires the consultation

[2] See http://bowland-files.lancs.ac.uk/corplang/lcmc/

[3] See http://bowland-files.lancs.ac.uk/corplang/ucla/

of the index of the Pearson's coefficient critical values set at different significant levels.

As a normal practice in corpus linguistics, we opt for the five percent as the threshold level to measure the strength of correlation between the two Chinese corpora. Given that we do not an obvious reason to assume or hypothesize the existence of a strong relationship between the two corpora in advance, we shall check the computed coefficient value with the critical value at the two-tailed non-directional category, which is always more prudent than using the one-tailed directional value.

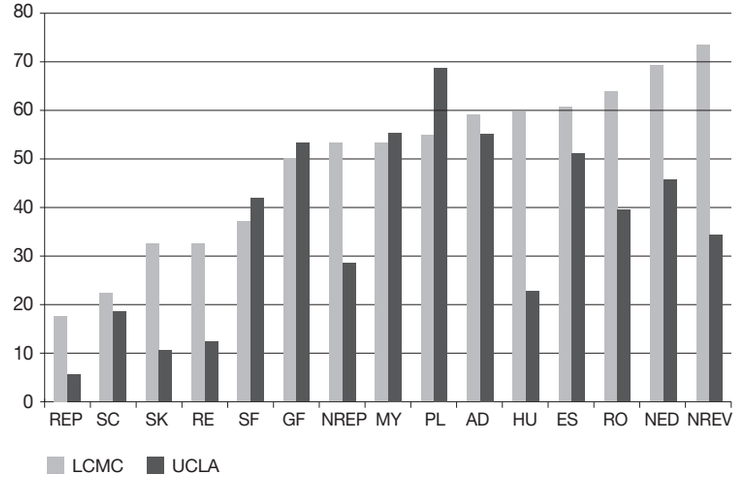
Table 2. Summary of Pearson's correlation test

| Statistic | Variable X (LCMC) | Variable Y (UCLA) |
|---|---|---|
| Mean | 317.8 | 180.7 |
| Covariance | 13364.4 | |
| Correlation | 0.4 | |
| Determination | 0.19 | |
| Degrees of Freedom | 13 | |
| Number of Observations | 15 | |
| Critical value for Pearson's test (two tailed at 5% level) | 0.5 | |
| Significance (Y/N) (two tailed at 5% level) | No (no significant correlation) | |

The mechanism of the Pearson's correlation test is that we start the statistical procedure by assuming a null hypothesis which treats the two corpora as having no relationship at all; and in order to subvert the default hypothesis, the computed coefficient must be equal or greater than the critical value. However, as Table 2 shows, the coefficient R obtained from the two Chinese corpora, which is as low as 0.435, is definitely below the threshold value at the critical five per cent, which is 0.514. The result suggests that despite the many similarities shared by the two corpora, such as the same sampling framework, the same language type, standard Mandarin Chinese, they indeed differ from



Figure 3. Comparison of normalized frequencies between LCMC and UCLA corpus




each other in terms of the frequency of occurrence and distribution of idioms.

To allow us to have an easier access to the numerical information provided in Table 2, the result has been used to draw a histogram in which the two coloured curves represent the distribution of idioms across different text types in the LCMC and the UCLA, respectively.

As may be seen from the graph, several important patterns regarding the evolving nature of Chinese idioms in written texts seem to emerge[4]. Firstly, bars with striped pattern which embodies the LCMC shows a general trend to run above bars with dotted lines, representing the UCLA corpus. This fits well with the descriptive statistics shown in Table 2, where the mean frequency of the LCMC is twice that of the UCLA corpus. This seems to suggest that at an overall level, the language recorded in the LCMC is more idiomatic than the material compiled in the UCLA, constructed some ten years later. However, idiomaticity is a complex concept which may well have different connotations in different text types or genres, for just as other languages, Chinese idioms or

[4] Both the LCMC and the UCLA corpus have been constructed with material collected from sources of written texts, such as online electronic libraries, or electronic texts posted on the www.

Cheng Yu as we call them in Chinese, tend to assume different aesthetic roles when applied in varying textual contexts.

As far as UCLA is concerned, the use of idioms is significantly lower than LCMC in six genres: REP, SK, RE, NREP, HU and NREV. These six textual genres are noticeable in the graph, due the sharp decrease in the use of idioms in the relevant text materials. It is interesting to see that these are invariably non-fictional Chinese text genres which represent a formal language register in accordance with the Chinese writing conventions, which is especially the case of NREP (news reportage), NREV (news reviews), RE (religious), REP (reports and official documents).

Such finding suggests that the pragmatic function of idioms in Chinese writing is gradually evolving towards an informal style, since its prominence in formal writings in classical Chinese appears to be diminishing in formal contemporary mandarin Chinese writings. On the other hand, the rhetorical or aesthetic value of idioms in Chinese fictional or popular writings has been steadily enhanced, which is well represented by the two small peaks along the pink line as above its blue counterpart: SF (science fiction), GF (general fiction) and PL (popular lore). The comparison of two large-scale monolingual Chinese corpora have thus provided useful contextual information which helps us to understand Liu's enhanced use of Chinese idioms in his translation of *Don Quijote* when compared to Yang's early version.

An important type of corpus tagging method worth further explanation is the proposed problem-oriented annotation (see Figure 1). By definition, problem-oriented annotation refers to a corpus annotation scheme which highlights and focuses on linguistic features that are most relevant to the research question to minimise the cost entailed by a full corpus annotation. This is a pragmatic annotation strategy which is particularly relevant to the construction of small-scale and topic-specific corpora.

The proposition of such an annotation method is due to the fact that at the moment, no such a "perfect" corpus encoding system exists which would achieve a one-hundred-percent precision rate when working on different corpora. As a result, more than often, the corpus information generated is a mixture of valid data and false data. The identification of valid data as an immediate solution will make up for the fallible nature of most current corpus tools. That is, one has to isolate valid linguistic information from an agglomeration of corpus data generated by automatic computational tools.



Table 3. Workflow of corpus-based translation studies

| STEP | MAIN RESEARCH TASKS |
|---|---|
| 1 | Text sampling and corpus construction |
| 2 | Corpus text pre-processing, e.g. segmentation, character code conversion, lemmatization, alignment (for translational corpora), etc. |
| 3 | Corpus annotation or marking-up, e.g. syntactic, part-of-speech, semantic, pragmatic, discoursal, problem-orientated, etc. |
| 4 | Corpus data retrieval and pattern recognition |
| 5 | Quantitative analysis and theoretical model construction |
| 6 | Testing the theoretical model on a larger set of corpus data |

## 3. Conclusion

There are two main criteria for the identification of valid corpus data in language corpora which are (1) corpus data must be easily machine-retrievable and (2) they must be suitable for quantitative corpus analysis. It should be noted that despite that some linguistic devices and categories represent important features of translational corpora, they may not be suitable candidates for the corpus-based translation analysis, as they either require sophisticated corpus techniques that are too expensive to develop, or not sufficient enough for a proper statistical analysis. Table 3 shows the workflow of a typical corpus-based translation project. The significance of the problem-oriented annotation, or in other words, the selection and analysis of corpus data satisfying these two conditions will greatly reduce the cost of extensive corpus annotation, and facilitate the identification and retrieval of useful textual and linguistic patterns in corpora in Step 4.

In conclusion, corpus-based translation studies from its inception in the 1990s remains largely dependent on the development of relevant digital language data bases and pragmatic research methodologies. There are three main issues that one has to bear in mind when pursuing corpus-oriented translation projects which are firstly, how to alleviate the labour-intensive nature of manual analysis; secondly, how to establish appropriate descriptive frameworks for systematic textual analysis; and thirdly, how to establish testable

hypotheses or replicable models that may reveal the nature of the texts under investigation. In other words, an original corpus-based translation study should aim to tackle at least one of three key research issues which are (1) development of effective corpus annotation techniques to enhance the balance between manual and automatic textual analysis; (2) testing of the systematicity and replicability of empirical analytical models to process corpus information; and (3) development of theoretical hypotheses to reveal the nature and character of translated texts.



## References


Baker, M. (1999) "The role of Corpora in investigating the linguistic behavior of professional translators", in *International Journal of Corpus Linguistics*, 4, pp.281-98

Baker, M. (995) "Corpora in Translation Studies: an overall view and some suggestions for future research", in *Target*, 7, pp. 223-43

Brown, P. F. et al. (1991) "Aligning sentences in parallel corpora", in the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pp. 169-76

Deng, Y. et al. (2006) "Segmentation and alignment of parallel text for statistical machine translation", in *Natural Language Processing*, Cambridge University Press, pp. 1-26

Gale, W. A. and Church, K. W. (1991) "A program for aligning sentences in bilingual corpora", in the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pp. 177-84

Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge University Press

Ji, M. (2010) *Phraseology in Corpus-Based Translation Studies*, Oxford: Peter Lang

Kenny, D. (2001) *Lexis and Creativity in Translation: A Corpus-Based Study*, Manchester: St. Jerome

Laviosa, S. (1997) "Investigating Simplification in an English Comparable Corpus of Newspaper Articles", in K. Klaudy and J. Kohn (eds.) Transferre Necesse Est. *Proceedings of the 2nd International Conference on Current Trends in Studies of Translation and Interpreting*, Budapest, Hungary. Budapest: Scholastica. pp. 531-540;

Laviosa, S. (2000) "Simplification in the language of translation: before and after the advent of corpora", *Athanor,* XI (3)

Lawler, J. M. & Dry, A. (1998) *Using Computers in Linguistics: A Practical Guide*, London: Routledge

Li, Yu Ling (2003) "Linguistic inheritance of idioms from ancient Chinese", in *Journal of Kai Feng University,* vol. 17, no.4, pp. 44-47

McEnery, T & Wilson, A (1993) "Corpora and Translation: Uses and Future Prospects", in http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf;

Mundy, J. (1998) 'A Computer-assisted Approach to the Analysis of Translations Shifts', in L' approche basée sur le corpus/ The Corpus-based approach, special issue of *Meta*, XLIII (4), guest edited by Sara Laviosa, pp. 542-556

Oakes, M. and Ji, M. (2012) Quantitative Methods for Corpus-Based Translation Studies, Amsterdam and Philadelphia: John Benjamins

Pápai, V. (2004) "Explicitation, a universal of translated texts?" in Mauranen & Kujamäki (eds.) *Translation Universals: Do They Exist?* Manchester: John Benjamins, pp. 143-65

Piao, S. (2002) "Word alignment in English-Chinese Parallel Corpora", in *Literary and Linguistic*

*Comptuing*, 17 (2), Oxford University Press, pp. 207-30
Puurtinen, T. (2004) "Explicitation of clausal relations: a corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature", in Mauranen and Kujamäki (eds.) *Translation Universals: Do They Exist?* John Benjamins, pp. 165-76
Saldanha, G. (2005) Style of Translation: An Exploration of Stylistic Patterns in the Translations of Margaret Jull Costa and Peter Bush, PhD dissertation, Dublish City University
Stubbs, M. (1996) *Text and Corpus Analysis*, Oxford: Blackwell
Teubert, Wolfgan (1996) "Comparable or Parallel Corpora?", in *International Journal of Lexicography*, vol.9, no.3, Oxford University Press, pp. 238-64
Thomas, J. and Short, M. (eds.) (1996) *Using Corpora for Language Research*, London: Longman;
Véronis, J. (2000) Parallel Texts Processing: Alignment and Use of Translation Corpora, Kluwer Academic Publishers
Xiang, G (1979) "Relationships between Chinese Idioms, Natural Environment, Cultural Traditions, and Linguistic Characteristics", in *Chinese Language*, vol. 2, pp. 112-121

---

***Author's email address***
meng.ji@uwa.edu.au

***About the author***
Meng Ji is Associate Professor of Translation Studies at the University of Western Australia. Her areas of research include corpus-based translation studies, contrastive linguistics (Chinese / Japanese / Spanish / Engli sh). She has published seven books and is the founding and current editor of Routledge Studies in Empirical Translation.